

An Exploratory Design Science Research on Troll Factories

Francisco S. Marcondes, José João Almeida and Paulo Novais

ALGORITMI Research Centre/LASI, University of Minho, Braga, Portugal

E-mail: francisco.marcondes@algoritmi.uminho.pt

Abstract. Private and military troll factories (facilities used to spread rumours in online social media) are currently proliferating around the world. By their very nature, they are obscure companies whose internal workings are largely unknown, apart from leaks to the press. They are even more concealed when it comes to their underlying technology. At least in a broad sense, it is believed that there are two main tasks performed by a troll factory: *sowing* and *spreading*. The first is to create and, more importantly, maintain a social network that can be used for the spreading task. It is then a wicked long-term activity, subject to all sorts of problems. As an attempt to make this perspective a little clearer, this paper uses exploratory design science research to produce artefacts that could be applied to online rumour spreading in social media. Then, as a hypothesis: *it is possible to design a fully automated social media agent capable of sowing a social network on microblogging platforms*. The expectation is that it will be possible to identify common opportunities and difficulties in the development of such tools, which in turn will allow an evaluation of the technology, but above all the level of automation of these facilities. The research is based on a general domain Twitter corpus with 4M+ tokens and on ChatGPT, and discusses both knowledge-based and deep learning approaches for smooth tweet generation. These explorations suggest that for the current, widespread and publicly available NLP technology, troll factories work like a call centre; *i.e.* humans assisted by more or less sophisticated computing tools (often called cyborgs).

Keywords: social media agent, troll factory, counter propaganda

1. Introduction

Propaganda and counter-propaganda have long been used by governments (including the military) and private companies. However, online social media provide a special environment for propaganda [1] that is being exploited by groups of individuals who spread propaganda and conduct cyber-attacks, often referred to as *troll factories*. As troll factories are obscure facilities, their inner workings and especially their technology are largely unknown.

Understanding troll factories is a cornerstone of prevention and countermeasures. Most attempts in this direction are based on: a) compiling information leaks to build a coherent picture [2]; b) observing the rumour chain in online social media to infer how they work [3]; or c) examining unclas-

sified military material [4]. This paper follows an alternative approach based on Design Science Research (DSR) [5, 6] but apply it to reverse engineering.

The idea of reverse engineering is to understand how a technology works [7]. If the artefact to be studied is available, a systematic study of it can be carried out. If it is not available, logic is used to infer the technology from images, descriptions and possibly parts of it. If almost no information is available, as is the case for this paper, a speculative design can be used to explore what the technology might be and the likely limitations it might have.

This paper has a fairly extensive introductory section, which aims to provide a broad context for the research that will be undertaken. First, a landscape of the spread of fake news is provided

to show the breadth and complexity of the issue. Then, as this paper is not based on a positivist stance, a longer explanation of the methodology is recommended. With the methodology explained, it is possible to draw the working hypothesis for this paper. This leads to the description of the setup in which the experiment will be conducted. The introduction ends with an ethical disclaimer.

The following section (section 2) presents the exploratory design experiments in which the possibility of building the artefacts discussed is estimated. Since it follows an inductive reasoning, two steps are considered; one discussing the base case (section 2.1) and another generalising inferences from the base case (section 2.2). Induction considers single and multiple bots interacting with single and multiple users. The following section (section 3) discusses the results of the experiments considering the proposed hypothesis. The last section presents some concluding remarks and future work.

Note that Twitter has recently been rebranded as X, but since the name Twitter is still widely known, it is retained for this paper.

1.1. Fake News Spreading Landscape

A *troll* is a person, often behind one or more inauthentic profiles, who posts or engages in inflammatory, disingenuous, and rambling actions on the Internet aimed at provoking emotional responses [8]. This is a convenient attitude for propaganda, which can be defined as «the communication designed to influence a belief or action, or to inculcate attitudes in a large audience, in a way that bypasses or suppresses an individual’s adequately informed, rational, reflective judgement» [9]. A *troll factory* is then an organised group of trolls working towards the same goal, nowadays often propaganda [10] or cognitive operations [11], in online social media [2]. Fake news, in turn, is an umbrella term referring to disinformation, malinformation and misinformation [12], currently often associated with propaganda on the Internet.

An online social media post containing fake news that reaches a wide audience is said to have gone *viral*. In a successful campaign, several posts containing fake news have gone viral, which is called *infodemics* [13] (term often related to, but not limited to, healthcare information). The virus metaphor is often deepened through the SEIR

model (SEIR stands for susceptible, exposed, infectious, recovered) but applied to online social networks [14]. In this sense, the work of a troll factory is to find susceptible people, expose them to propaganda, and make them infectious. The work of counter-propaganda is to prevent this procedure from happening and to turn infectious people into recovered.

Online social media is a conducive environment for infodemics. There are two key features of online social media that are the cornerstone of this effort. One is that these platforms are capable of triggering the dopamine reward behaviour training [15], which acts as a *Skinner box* [16]. Another is the mediation algorithm which, in addition to reinforcing behavioural training and yielding to various cognitive biases due to segmentation [17], enables the targeting of susceptible people to be sent the malicious content [18].

In order to exploit the features as such, a large number (eventually millions or tens of millions of [19]) of inauthentic profiles is required [2]. This group includes bots, cyborgs, paid followers, *etc.* In the bot group there are several types such as influencing bot, echoing bot, spamming bot, *etc.* [20]. Twitter estimates that 5% of its accounts are bots ($\approx 48M$), while an independent institute claims it could be as high as 20% [21]. Nevertheless, it is not always about inauthentic profiles, there is a critical mass situation that occurs when there is a sufficient number of radicalised people that makes the use of the bots unnecessary.

A requirement to be met by the mediation algorithm is to provide an infinite timeline [22] to keep people online as long as possible [18]. Consider a mediation algorithm that uses a semantic network to relate someone’s interests. Let this person be tagged as interested in astronomy. After exhausting the recent content on astronomy, which is not expected to be much, the mediation algorithm might start showing posts on astrology, flat Earth (Earth is an “astro”), *etc.*; see figure 1.

The infinite timeline would facilitate the spread of fake news (and perhaps radicalisation) [24] as long as it exists in quantity. This feature is related to the agenda setting phenomenon [25], *i.e.* the amplification of content by exposing it to a wide audience. The more people talk about something (whether in agreement or not), the greater the change in a susceptible person exposed to a particular speech. It also triggers the group con-

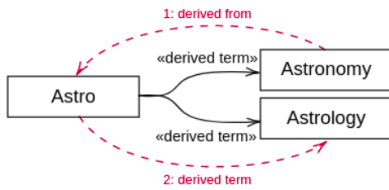


Fig. 1. ConceptNet [23] shortest path from astronomy to astrology

sensus effect [26], in short, if “everyone” in a community is committed to an idea, there is a tendency for the dissonant individual to merge with the crowd.

Finally, at present, it is not possible to say that spreading rumours is an activity limited to a single online social media. For example, some malicious content on COVID-19 started on 4chan, went through Telegram and Gab, and finally reached Facebook [27]. It is not the case that Facebook is safer than the other media; it seems to be the best place to spread links [28]. Thus, each online social media is used for different purposes and probably for different audiences [29].

Microblogging (*e.g.* Twitter, Gab, GETTR, TruthSocial, *etc.*) plays an important role in this mosaic. Twitter, or rather microblogging, seems to be the preferred platform for many leaders associated with the alt-right (*e.g.* Donald Trump, Jair Bolsonaro, *etc.*). It has not been possible to find an in-depth explanation for such a preference, and it is beyond the scope of this paper to discuss it. Nevertheless, it is known that Twitter is a primary recruitment source used by terrorist groups [30]. It can therefore be suggested that microblogging, more than a broadcast channel, is perhaps a lynchpin in the recruitment of supporters.

For remarks, note that not all troll factories are professional (*i.e.* skilled in exploiting online social media features) or propaganda-oriented. For example, Desinfomedia is a company that publishes an online tabloid aimed at making money through Google AdSense by spreading rumour [31]. In another example, a group of trolls sent strobing images on Twitter targeting people suffering from stroboscopic epilepsy [32].

It is also worth noting the difference between propaganda and marketing. In a nutshell, the ultimate goal of propaganda is to spread an idea [33] whereas that of marketing is to sell products [34]. An overlap between the two is inevitable, but be-

cause the goal is different, so must the strategy. For example, social media propaganda is often based on [a swarm of] inauthentic accounts (*e.g.* bots, cyborgs and trolls) [2, 3] organised into a botnet as coordinated agents *cf.* [35], whereas social media marketing must rely on an authentic corporate account and advertisers. It is therefore not easy to use marketing strategies to spread propaganda.

Note also that there are slight differences between propaganda, counter-propaganda and civil propaganda; a propaganda operation is usually an attack from a foreign country (warfare) or a competing company (dirty strategy), so the source of the propaganda must be kept hidden. For counter-propaganda, a credited source backed by an echoing botnet is often a suitable setup [33]. Civilian propaganda, on the other hand, usually has a credited source, *e.g.* the president of a country, spreading or reinforcing a propaganda [36].

Also, not all bots are designed for propaganda. In social media marketing, chatbots are used to provide highly available customer channels [34]. There are also bots used to spread religious messages [37] and a botnet of over 350k bots randomly quoting the Star Wars novel without any other apparent behaviour [38].

In summary, the spread of fake news is a complex “ecosystem” composed of different actors distributed across different environments.

Research question. One question that arises from this understanding is on the nature of trolls. Are they humans or bots? In other words, can a troll factory be a fully automated facility? Answering this question is essential for assessing the threat level and planning countermeasures. The aim of this paper is to provide such an answer.

1.2. Methodology

Design science, also known as the science of the artificial, is a research method proposed in [40] for dealing with technology. In short, technology research differs from science in the sense that technology is concerned with producing artefacts, whereas science is concerned with understanding phenomena. The difference between applied science and technology is that the former relies on science, whereas the latter does not [41, 42]. Simply put, historically, technology has progressed without the need for science; for example, primitive

Table 1
Comparison between three philosophical stances [39].

Paradigm	Positivism	Interpretativism	Constructivism
Ontology	One unique reality, knowable and probabilistic	Multiple realities, socially constructed	Many alternatives, contextualised in different environments Socio-technological
Epistemology	Objective, unbiased search for truth	Subjective, knowledge emerges from interaction between researcher and subject	Knowledge built from design choices, concrete artefact building within a given context. Iterative definitions of reality.
Methodology	Observation, quantitative, statistical	Participatory, Qualitative, Hermeneutic, Dialectical	Development, measuring the impact of the artefact within its environment.
Axiology	Universal and predictable truth	Descriptive and contextualised understanding	Creation/Extension; Understanding

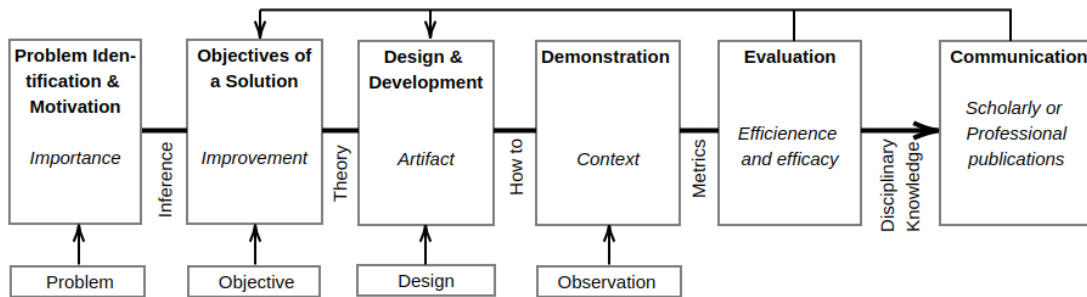


Fig. 2. The steps of the Design Science method *cf.* [5]

people developed bows and arrows without understanding the underlying mechanics.

It is beyond the scope of this paper to present a discussion between the various existing research paradigms. Table 1 summarises the main differences between constructivism (the underlying Design Science paradigm) and other common research paradigms. It is worth mentioning that the experiments carried out on the constructivist paradigm (see methodology in table 1) are based on the production of artefacts (either an artefact can be built or not) and on the measurement of its impact in the *world of the problem* (to what extent an artefact is able to change the environment to the desired state) *cf.* [43]. Note that there are several levels of development that an artefact can achieve, see [44] for a reference. In short, the development level of an artefact can be summarised as proof of concept, prototype, or product; expected to run in laboratory, relevant and operational environments, respectively. The Design Science method is shown in figure 2.

Perhaps the most appropriate approach to the problem of this paper is *reverse engineering* [7]. In short, reverse engineering uses deductive reason-

ing to understand, or at least gain an insight into how a technology works based on the information available. It should be emphasised that it is not always possible to have access to the artefact being reversed [45], in which case the most likely solution is considered reasonable. For example, during the Second World War, some enemy technologies were reversed by inferring from field information such as descriptions and photographs.

The case for this paper is a little more difficult, because if the troll factories are fully automated, the number of managers would be so small that no leakage would be expected. On the other hand, if the trolls are humans, not all trolls are expected to be above user level, so the descriptions would be of little use. Finally, assuming the existence of a technical team, it would be expected that these professionals, with the exception of a few architects, would be so compartmentalised that they would not be able to understand the big picture; any leakage would therefore be incidental. Since there is no material for inference, the alternative is to carry out an exploration *cf.* [46]. Exploratory programming is an iterative and incremental middle-out design approach in which disposable artefacts

are elaborated just enough to provide the understanding needed to carry out the inference.

1.3. Design Hypothesis

The falsifiability principle states that the cases that are more likely to falsify the hypothesis must be pursued [47]. Applied to the context of this paper, the design hypothesis must be that it is possible to build a fully automated troll factory. Therefore, research must focus on the artefacts that are central to the design, but also less likely to be possible to build; *i.e.* the most wicked essential feature. A wicked feature is a feature whose properties correspond to those of wicked problems *cf.* [48, 49].

Roughly, a troll factory has two main activities, *sowing* and *spreading* [2], which are intertwined [50, 51]. The sowing activity aims to create and maintain a social network that can be used for the spreading activity. In short, it is the same profile that engages in social interactions with its network to opportunistically spread malicious content. As an example of the alternation between these two stages, it is not uncommon for fake profiles to be presented with a photo of a young girl in a bathing suit, which acts as “bait”, and then change to a middle-aged man in a suit for the spreading stage.

One task in the sowing activity is to produce *smooth posts*. A smooth post is a tweet that the reader would find worthy (*e.g.* interesting, funny, *etc.*) and that leads to engagement (*e.g.* to a `like` action). In other words, it is a post whose content is in line with the reader’s beliefs, desires, and values [17]. A *smooth interaction* occurs when a smooth post results in a positive response; *i.e.* when the actors reinforce each other’s attempts to share information [52].

Thus, it can be said that sowing is more wicked than spreading. In short, the sowing task is a long-term relationship-building activity, subject to all sorts of problems that can come either from the online social media or from the actors [53]. Considering the possible interactions that can be carried out in an online social media, producing a smooth content for a post is certainly more wicked than the other types of interactions such as `like` and `share` [54]. In terms of text, there are at least two types of content [55], *news post* (consisting of a link with or without a snippet of text) and *opinion post* (expressing a speech act, see [56]). The sec-

ond is more wicked, as it requires coherence with a fact, news, or post. In this sense, a replying post is more wicked than an initial post.

The automatic production of smooth replying content for a single bot is bad enough. This is because it is not just a matter of producing content for a single reply, but, as suggested, for several posts over a long period of time, in such a way that some coherence is maintained, including the bot’s feed. For example, the exact same post should not appear more than once in the feed (except in special circumstances). However, a troll factory operates tens of thousands of bots within a botnet. So it is not just one bot’s feed, but several. For example, it would be awkward if two or more profiles participating in someone’s network posted the exact same phrase (again, except in special circumstances). So such a tool needs to work in a way that avoids the *uncanny valley* *cf.* [57].

Furthermore, if it is true that microblogging platforms are used for recruitment, it can be assumed that this is a suitable place to sowing attacks. It is therefore a suitable target environment. This is opportune because microblogging is perhaps more wicked than the other text-based social media platforms due to its reduced post length that makes it easier for followers of a profile to realize when a post is being repeated, does not make sense, *etc.*

Working hypothesis. Finally, the working hypothesis for this paper is that it is possible to design an artefact capable of automatically producing smooth replying post content for a botnet running on a microblogging environment.

A caution with the exploration of such hypothesis is that it must not fall into imagination. Thus, any technology that is not publicly available or cannot be built, cannot be considered for the experiment. In addition, considering average private troll factories, they are not expected to use top of the hedge or obscure NLP technology to support their activities due to the cost of implementation and maintenance. Therefore, only widespread and publicly available NLP technology is considered for the experiment. The exploration is carried out on a Fujitsu LIFEBOOK A3510 notebook running Ubuntu 22.04.1 LTS and Python 3.10. The main external modules are `CMUTweetTagger`, `emoji`, `nlTK`, `numpy`, `pandas`, `textblob`, and `twint`; organised as the figure 3.

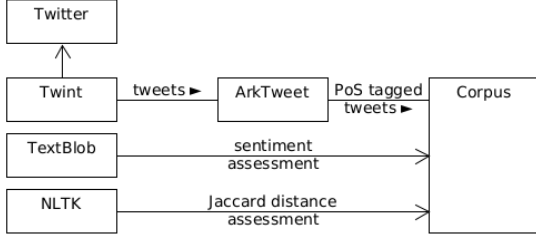


Fig. 3. Diagram showing the main data flow and connections between the main modules used.

1.4. Smooth Tweet Construction by Mirroring

One question that arises in addressing this hypothesis is what would be the design for an artefact as such, this is answered in [53, 58]. Another question is whether it is possible to generate automatic post content regardless of whether it is smooth or not, the answer if positive is discussed in [54, 55]. This paper can then be seen as an extension of those four.

Microblogging platforms such as Twitter can be understood as a kind of structured dialogue *cf.* [59], as

$$f : \mathbb{T} \rightarrow \{\mathbb{T} \mid \emptyset\} \quad (1)$$

where \mathbb{T} is the set of all tweets. So a Twitter dialogue consists of a tweet followed by a tweet or none chained as a recursive tree [53]

$$\begin{array}{c} \text{---} t \text{---} \\ \text{---} \text{---} \text{---} \\ | \\ [t' = t]_0 \quad \dots \quad [t'' = t]_n \end{array} \quad (2)$$

The concatenation is given by equating t and t' , *e.g.* $t \rightarrow [t', t] \rightarrow [t'', t] \rightarrow \dots, \forall t \in \mathbb{T}$. Note that each tweet is close to a context-free statement.

In this sense, it is possible to establish an attribute expression *cf.* [60] to keep the semantic coherence [61, 62] between posts [53] as

$$t_\gamma \Rightarrow t'_\gamma \quad (3)$$

where γ can be any property of interest, such as a sentiment, a personality trait, a social norm, *etc.* [53]. For instance, replying «I don't like basketball» for a tweet such as «My kid won his basket-

ball game!!» is not semantically coherent at the sentiment dimension; it is adversarial [63].

The γ applies to any tweet feature [64, 65]. Let the elements of a tweet t be mentions m , snippets s , text x , emojis e , link l , hashtags h and rich media r , see figure 4 for a reference, then

$$t = msxelhr \quad (4)$$

$$m_\gamma s_\gamma x_\gamma e_\gamma l_\gamma h_\gamma r_\gamma \Rightarrow m'_\gamma s'_\gamma x'_\gamma e'_\gamma l'_\gamma h'_\gamma r'_\gamma \quad (5)$$

A structure as such is useful for building rapport by reproducing someone's beliefs and feelings about something [3]. Establishing rapport is important for building trust between peers, so it is an essential feature of sowing. This approach is also known as *mirroring*. The rapport is established when one's ideas are recognised in another, which in turn requires time and a sufficient number of tweets.

The γ can be handled in several ways, from item-wise, where each element in the base tweet maps to the corresponding element in the replying tweet, to n-dimensional vector structures, perhaps approached by cosine similarity. Another possibility is to use averages, which might include information about the tweet being replied to (\bar{t}_γ) as well as information about the profile posting the tweet or the target audience (\bar{a}_γ). Note also that there could be multiple γ evaluated for the same tweet as $\bar{t}_{\gamma, \dots, \gamma'}$ and audience as $\bar{a}_{\gamma, \dots, \gamma'}$. To illustrate, this could be expressed as

$$\bar{t}_{\gamma, \dots, \gamma'} = \frac{\sum_t t_{\gamma, \dots, \gamma'}}{l_t} \quad \text{and} \quad \bar{a}_{\gamma, \dots, \gamma'} = \frac{\sum_a \bar{t}_{\gamma, \dots, \gamma'}}{l_a} \quad (6)$$

1.5. Ethics Statement

Given that social media interactions affect people's psychological states, researching them is like

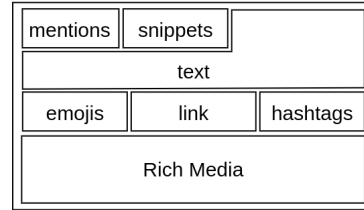


Fig. 4. Representation of elements in a tweet.

researching people. Therefore, the authors felt it was important to emphasise that no experiment was conducted in actual online social media. However, this did not affect the results of this paper, as due to its current maturity *cf.* [44], running the prototype in an operational environment was never expected. In fact, the results obtained suggest that this was not the case. All data collected was publicly available at the collection stage and de-identified *cf.* [66] for privacy reasons [67].

A paper like this could be accused of developing technology for use in propaganda. In response, this is an individual initiative, the scale of which cannot be compared to fully funded development teams. Furthermore, its exploratory nature precludes the production of fully developed artefacts that could be used to cause harm. However, even in the unlikely event of a breakthrough, the fact that it is in the public domain would help people to understand and counter any propaganda that could arise from it.

2. Design Experiment & Results

Since the falsifiability principle is applied, research has so far followed a hypothetical-deductive rationale, searching for the most wicked property to build.

Once identified, it is necessary to provide a constructivist proof, *i.e.* to build the intended artefact, or to show that it is not possible to build it (at least with current technology). An inductive procedure is then proposed as an appropriate reverse engineering approach. Note that the inductive reasoning used for the experiment is not opposed to the deductive reasoning used for hypothesising, as they are independent steps in the design research method. Induction here means exploring possible designs for a concrete scenario as a base case, and then discussing the suitability of these proposals for constructing an artefact that would fit the general case. The hypothesis is evaluated at the induction stage. Also, since sowing is an open domain situation, it is not worth considering the most wicked case, as it is likely to be a marginal case; average cases are preferred.

On addressing language-intensive applications, there are currently two paradigms to consider [68]: *corpus linguistics* and *language models*. In addition to these paradigms, the *naive NLP* paradigm,

which is mostly based on heuristics, can be included. These three paradigms are treated as dimensions in this exploration. Note also that the design exploration follows an iterative and incremental approach, *i.e.* the simplest design that delivers value is considered and complexity is added incrementally.

For an organization, this section is divided into two sections, one dealing with the base case (section 2.1) and one dealing with the induction step (section 2.2). Thus, sections 2.1.1, 2.1.2, and 2.1.3 discuss the base case of the *Naive NLP*, *Corpus-based*, and *LLM-based* approaches, respectively. In the induction part, different setups are proposed, split between sections 2.2.1 and 2.2.5.

2.1. Base-case

For the case study, consider the tweet in figure 5a. This is a random tweet with no particular property. For reference, figure 5b shows the reply the tweet received at the time of collection. Thus, figure 5 shows a typical random Twitter interaction serving as discussion basis for this exploration.



Fig. 5. A typical Twitter interaction (de-identified).


2.1.1. Naive NLP approach

There are several NLP approaches that can be used to generate replies to tweets [68, 69]. Roughly, it consists of parsing the text to extract features of interest and building the reply based on them. This approach is called naive in the sense that it



overlooks actual features present in most tweets and relies on heuristics.



Following the expression 5, perhaps the simplest solution would be an element-wise match along a dimension. So let the dimension be sentiment valence and text subjectivity, then $x_{v,s}$, which is assessed by the TextBlob's [70]. For the text in the figure 5a, subjectivity is $s \approx .57$ and valence is $v \approx .44$. Thus, considering the interval ± 1 , a set of semantically coherent words was retrieved from a lexicon of common words by the formula $| (w_s^n - x_s) + (w_p^n - x_p) |$ for producing $x_{v,s} \rightarrow x'_{v,s}$. The best matching words are **cool, healthy, interesting, safe, sincere, sweet, true**. Thus, for an instance,

sweet (i)

A similar approach would be to consider cross-element productions, aiming to get closer to the case study, it would be $x \rightarrow e'$. Despite the existence of lexicons that relate emoji to sentiment valence, such as [71], which would probably succeed in reproducing the case study, the alternative chosen for this example is to search for emoji based on the words in the text. The words that have matching emoji from a list of emojis are **heart and movie**; some of the related emojis are: . Thus, for an instance,

 (ii)



In both examples, as can be seen in figure 5b, the solution produced would be suitable as a smooth reply; although some of them are a little awkward, it should be noted that some awkwardness is also a feature of Twitter posts. The next iteration is straightforward and would be to combine these two approaches to create an increasingly featured response such as “*sweet* ”. This can be further improved by incorporating heuristics such as extending the vowel *cf.* [72], *e.g.* “*sweeeet* 

sweeeet   #LOVE #MOVIE (iii)

The next natural step is to consider the use of n-grams, which would eventually reach the point

of producing reply templates. At this point, the exploration requires a paradigm shift, leading to a fully featured case-based reasoning system based on either AIML (Artificial Intelligence Markup Language) [73] or Rasa [74]. Although these tools are expected to handle multiple cases, recall that this discussion is still considering the base case presented in figure 5.

The AIML alternative is based on matching an input to a pattern and retrieving a template. The template part is straightforward in the sense that it can be programmed to be enriched with the smooth function discussed so far; the pattern part is not so straightforward. There are two strategies to consider: one is to create a very general template based on the **Netflix** and **movie**. Another would be to create a template for each movie or series in the Netflix. The first template would be necessarily general and thus likely to fall into the uncanny valley, whereas the second would be quite convincing at the cost of ongoing maintenance. Considering the second case, it could return, after reinforcement, something like

I love the way it shows the ups and downs of the two perspectives. The actress who plays (iv) Natalie, WOW!   #LOVE #MOVIE

The Rasa alternative is based on probabilistic intent detection that retrieves a deterministic response. So if used for the training part, this tweet text would be included as an example of, say, the *uncompromised interaction* intent that would call the smooth function described so far to return the response. Because Rasa is based on Transformers technology, it is able to generalise to any post that is similar to the example. Despite this advantage, Rasa does require some fine-tuning and recommends special hardware to run properly. Although it would probably be best to fit the tweet into a pattern, the response would be quite the same as the one presented in quote iv.

In terms of results, balancing the strengths and weaknesses of both approaches, it is possible to suggest that for the average case they are equivalent and would provide sufficient performance [75]. Consider also that Twitter, apart from the direct messaging feature, has a type of dialogue in which it is more difficult to enter the uncanny valley compared to a direct dialogue [59]. It is then possible to suggest that any of these approaches would provide a convincing smooth reply to the case study.

2.1.2. Corpus-based Approach

The section 2.1.1 presented four productions that could be used to reply to the base case tweet. The problem with this solution is that it does not take into account how tweets actually are. For example, consider quote (iii); although it appears to be a common tweet, it is not. For example, it is not reasonable to assume that a real person browsing Twitter would be concerned about including hashtags in a tweet as such. In fact, only about 6% of the tweets use hashtags [55]. Furthermore, in the same example, the word *sweet* was extracted from a list of common English words; the question is whether or not the same list applies to Twitter. For answering questions as such, a corpus-based approach is recommended.

In short, a corpus is a dataset annotated with linguistic features. A major challenge in creating a general purpose corpus is to collect a collection of text that is sufficiently diverse to be considered representative of the language as it is spoken [76]. This is particularly difficult for Twitter [77, 78], as the collection strategy must avoid using the search engine and recursive crawling to ensure the randomness of the sample. For reference, most of the Twitter corpora on Kaggle are domain specific due to such difficulty.

Broad purpose corpora can be either general or monitoring. The first type, such as the Broad Twitter Corpus (165,739 tokens [77]) and Tweep-Bank (55,607 tokens [78]), are often too small. The second type, such as the Spritzer (2Gb daily files over ten years), is usually too large. For the purposes of this paper, a larger general Twitter corpus needs to be created. It is worth noting that, except for the monitoring corpora, each corpus is a snapshot of the language at a particular time and possibly place (*e.g.* before 2022 there was no King Charles III of England). Note also that it is not currently possible to perform independent scraping in Twitter [79], so the extractions are limited and potentially biased by the Twitter API.

To avoid these biases, a corpus was built using an onomatological approach, as described in the appendix A [1]. In short, this approach avoids common biases by randomly generating Twitter handles [80] and discarding the tweets from these accounts. Thus, the search is neither based on the search engine nor on crawling users' networks, which results in higher randomness. The resulting corpus is composed of 4,412,656 tokens from

253,720 English tweets produced by 2172 different profiles that were active at the time of collection. It is therefore sufficiently representative to provide a snapshot of the language. An overview of the results can be found in the diagram 6, the glossary of tags in the table 2.

On the basis of this corpus, it is possible to continue with the exploration of smooth tweets. Repeat the experiment that generated the phrase (i), but based on the corpus, it retrieves the best matching words that are actually used in Twitter (ordered by frequency): *secure, advanced, bizarre, incomparable, risk-free, non-violent, securely, wealthy, ecological*. Again, perhaps with the exception of the word *bizarre*, each could be considered a smooth Twitter reply.

Thanks to the corpus, it is now possible to check whether or not an answer consisting of adjectives like the ones presented is a common structure on Twitter. When querying the corpus, we find that tweets consisting of a single adjective are the 53th most common structure on Twitter (percentile 0.9995). Adding a punctuation mark makes it the 43th most common structure (same percentile), and adding a mention makes it the 7th most common structure (percentile 0.99997). So

@pto secure. (v)
is a very common Twitter structure. Note that this is a counterintuitive result, as you would expect an answer like (iv) to be better than (i) or (ii), but it is not. This actually makes sense considering the actual use of Twitter, people do not have the time to produce elaborate texts like (iv), especially for each subject.

For reference, table 3 shows the 10 most common syntactic patterns in the corpus that were used to automatically compose responses based on sentiment proximity and frequency. Tags that could not be associated with a sentiment, such as pronouns and determiners, for which only frequency is considered. Also, tags such as mentions, URLs and numbers are kept as placeholders. Again, any instance in the 3 table could be used for an uncompromised answer, but not always.

A closer look at the data shows that $\approx 97\%$ of the structures are unique, while the highest pattern represents 0.2% of the sample. The top 10 patterns shown in the table 3 represent 1.3% of the corpus. This suggests that not having a common syntactic structure is the most common pattern.

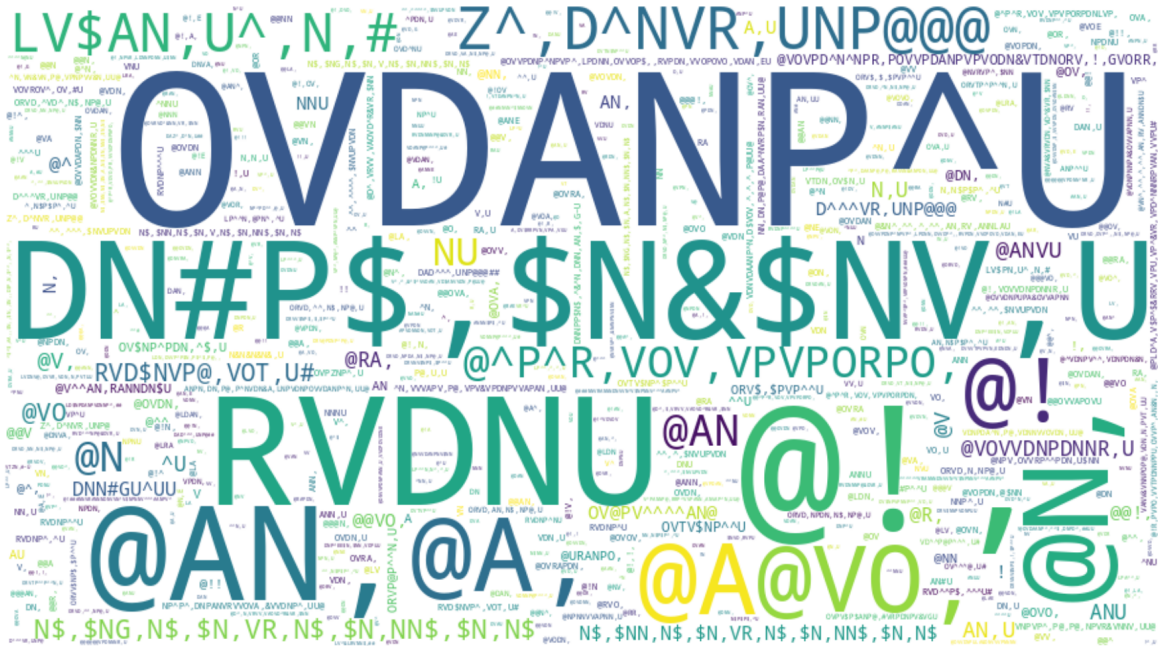


Fig. 6. World Cloud illustrating common tweet structures. See the tags description in table 2.

Table 2
Ark’s PoS Tagger symbols [72].

Tag	Description	Examples
N	common noun (NN, NNS)	books someone
O	pronoun (personal/WH; not possessive; PRP, WP)	it you u meeee
S	nominal + possessive	books’ someone’s
^	proper noun (NNP, NNPS)	lebron usa iPad
Z	proper noun + possessive	America’s
L	nominal + verbal	he’s book’ll iono
M	proper noun + verbal	Mark’ll
V	verb incl. copula, auxiliaries (V*, MD)	might gonna ought couldn’t is eats
A	adjective (J*)	good fav lil
R	adverb (R*, WRB)	2 (i.e., too)
!	interjection (UH)	lol haha FTW yea
D	determiner (WDT, DT, WP\$, PRP\$)	the eht its it’s
P	pre- or postposition, or subordinating conjunction (IN, TO)	while to for 2 (i.e., to) 4 (i.e., for)
&	coordinating conjunction (CC)	and n & + BUT
T	verb particle (RP)	out off Up UP
X	existential <i>there</i> , predeterminers (EX, PDT)	both
Y	X + verbal	there’s all’s
#	hashtag (indicates topic/category for tweet)	#acl
@	at-mention (indicates another user as a recipient of a tweet)	@BarackObama
-	discourse marker, indications of continuation of a message across multiple tweets	RT and : in retweet construction RT @user : hello
U	URL or email address	http://bit.ly/xyz
E	emoticon	:-) :b (: <3 o_o
\$	numeral (CD)	2010 four 9:30
,	punctuation <i>#, \$, ", (,), ., :</i>	!!! ... ???
G	other abbreviations, foreign words, possessive endings, symbols, garbage (FW, POS, SYM, LS)	ily (I love you) why (what about you) ’s -> awesome...I’m

Note that the difference between patterns includes repeating tags, *e.g.* $A \neq AA$. Applying Jacard distance $(|A \cup B| + |A \cap B| / |A \cup B|)$, where A and B are sets [68] for group equivalent patterns (*e.g.* with distance equal to zero), the unique patterns drop to about 10%. The problem with this approach is that syntactic context is lost, making rule-based generation difficult.

Table 3
Top-10 PoS patterns in the corpus.

#	Patterns	Instance
0	OVDANP^U	I enjoy this cool love in love <url>
1	@!,	<mention> okay.
2	RVDNU	most touching with love <url>
3	DN#P\$, \$N&\$NV,U	The top #love with <num>. <num> love and <num> top enjoying. <url>
4	@AN,	<mention> special #love.
5	@N,	<mention> #love
6	@A,	<mention> special!
7	@A	<mention> special
8	@VO,	<mention> enjoying this!
9	@!	<mention> ok

The natural next attempt would be to look for relationships between the words. However, the absence of syntactic patterns suggests the absence of n-gram structures. It may therefore be worthwhile to focus on language models.

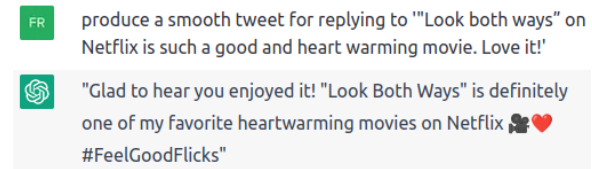
2.1.3. LLM-based Approach

An alternative approach to those presented here is that of large language models (LLMs) *cf.* [68]. In short, a language model takes a corpus as a data set and computes the relationship between the words in it. The currently most popular architecture for generating language models is Transformers [81], which uses the *attention mechanism* to score the importance of words in a sentence, thus improving context detection. A large language model receives as dataset different types of corpora from several sources. It is therefore useful to evaluate its ability to generate smooth tweets. It is worth noting that the computing power of Transformers of a context-free language [82].

The use of AI to automate tweet generation is not new, as it is a widely used digital marketing tool. There are several approaches, which, in addition to using a general Twitter dataset, include generating tweets based on the target user's own content [83]; interest or topic oriented [84]; persuasion tailored [85]; *etc.* However, most of these approaches are designed to produce targeted initial tweets, not replies, so most of the strategies are not directly applicable. In addition, there are models fine-tuned for Twitter in the Huggingface, but none of them stand out (the most trending model was downloaded 24 times in mid-2023) or have performance worth discussing. Then, using the existing approaches on the corpus created for this purpose is also unlikely to result in a model with good performance. This is left to the general purpose LLM to be assessed.

Perhaps the two best known LLMs currently available are ChatGPT and Bard. Figure 7 shows the response generated by these tools for the base case tweet in figure 5a. It is worth mentioning malicious LLMs such as WormGPT as it is claimed to be suitable for social engineering. As WormGPT is a paid Dark Web tool, it was not possible to evaluate it in the same way as ChatGPT and Bard. However, as suggested in [86], there is no particular concern regarding these tools as the performance is at most equivalent to any other LLM. The main difference between benign and malicious LLM is that the latter has no content moderation filters [86]. So malicious LLMs can be asked to produce malicious content (*e.g.* hate speech, phishing material, *etc.*) that is prevented on benign LLMs. For reference, unfiltered LLMs come closer to the Tay chatbot [87]. The point is that there is no

reason or evidence to support the idea that malicious LLM would be able to produce replies whose quality would exceed the results of benign LLM. Finally, it is possible to suggest that LLMs with similar training size are equivalent in practice.



(a) ChatGPT

✦ Sure, here are some smooth tweets you can reply with:

- "I totally agree! Look Both Ways is such a heartwarming and inspiring movie. I loved it too!"
- "Just finished watching Look Both Ways and I'm in love! Such a feel-good movie with a great message."
- "If you're looking for a heartwarming and uplifting movie, I highly recommend Look Both Ways on Netflix. You won't be disappointed!"
- "Look Both Ways is the perfect movie to watch on a cold winter night. It's heartwarming, funny, and inspiring all at the same time."
- "I'm so glad you enjoyed Look Both Ways! It's one of my favorite movies on Netflix right now."

You can also add a personal touch to your tweet by sharing why you liked the movie or what message you took away from it. For example, you could say something like:

(b) Bard (excerpt)

Fig. 7. Two LLM generations for replying figure 5a

Considering the responses shown in figure 7 and the actual response in figure 5, it is possible to suggest that for this particular case, there is no material adequacy in the generation. Furthermore, considering table 3, it is possible to suggest that these responses do not share a common tweet structure either. Nonetheless, each of these instances would be suitable smooth replies for the base case.

Bearing in mind that LLMs are short-term learners, it is often preferable to use prompt engineering *cf.* [88] than to go through a fine-tuning process [89]. A canonical prompt consists of: a) role; b) instruction; c) input data; d) context; and e) few-shot examples. Figure 8 shows the result of using this structure to generate the response. Note that the role and instruction would be the same

You are Twitter user. Produce a smooth reply to a follower's tweet

"Look both ways" on Netflix is such a good and heart warming movie. Love it!

On the eve of her college graduation, Natalie's life diverges into parallel realities: one in which she becomes pregnant and remains in her hometown to raise her child and another in which she moves to LA to pursue her dream career.

You just gave me #FridayNight watch. Thanks T.

See @Tutunuu

Who is watching life lessons on Netflix. When there is friends

=)



Absolutely loved "Look both ways" too! 🌟👏 It's such a touching exploration of life's choices and the paths they lead us on. Enjoy your Friday night movie night, and remember, there's always room for both heartwarming movies and good ol' "Friends" re-runs! 📺 #BalancelKey

Fig. 8. Engineered replying prompt (ChatGPT).

for each tweet, the input data is the tweet to reply to, the context was retrieved from the IMDb website, and the examples are actual replies to the tweets in figure 10. This is a prompt that could be generated automatically.

Note, however, that the response has both strengths and weaknesses. The strengths come from the context, which has enabled the model to generate a properly contextualised response. The weaknesses come from the response examples. The authors were not able to generate a text that followed the example texts (on ChatGPT 3.5); all generations were complete, well-written texts. Note that not all replies are smooth, resulting in awkward places in the generated text. So it would be necessary to filter the replies that can be used or to create a set of smooth examples to be considered on each case.

Finally, looking at the figure 7b, it is possible to see that, despite some variation, all generations follow a "style" of writing. This is the underlying premise of models used to judge whether or not a text is machine-generated, such as GPTZero (a tool for evaluating whether a text was written by ChatGPT), by training it on text generated by a given model, see figure 9. The point is that neither of these examples fits the common structures of tweets as shown in the table 3. So sooner or later it would enter the uncanny valley or be detected. An hybrid approach is a possibility to consider [90].

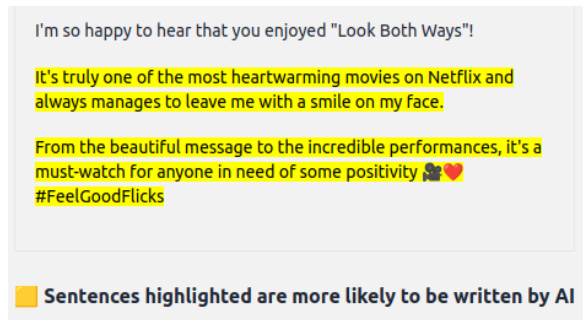


Fig. 9. GPTZero assessment for a ChatGPT generation. Since GPTZero requires a minimum of 250 characters, the ChatGPT was asked for rewriting the tweet for 280 characters (the maximum allowed length in Twitter).

2.2. Induction step

Continuing with the case study, figure 10 shows other random tweets on the same topic as the tweet shown in figure 5. Note that this is a regular induction step, since for a strong induction it would be necessary to consider every tweet. Regular induction is then the natural next step in this exploration.

The tweets shown in figure 10 are illustrations of the general case (not to be taken as individual cases), tweets that it would be expected to be handled by a topic specific replying tool.

Note that there are several other tweets that use the expression *look both ways*, for example in the context of car traffic, or even others, but are not included because they are considered off-topic for this induction.

As an estimate of the number of tweets on the subject of this case study, a search on Goggle using the query `site:twitter.com "look both ways" netflix` returns around 1200 results, almost all of them from mid-2022, when the film was released. A first issue appears here. Note that the number of tweets on this topic is extremely small and the time frame in which this topic arose is extremely short. It is then impractical to create rules, analyse patterns or fine-tune models for each specific topic as suggested in (iv). The actual procedure must then be a general solution or the involvement of humans. Only non-specific approaches are then considered in the following explorations.

For directing the discussion, as already mentioned, the dimensions considered are: a) a single bot handling multiple posts from multiple people; b) a single bot handling multiple posts from



Fig. 10. A sample of related tweets for illustration (de-identified).

a single person; c) multiple bots replying to a single post; d) multiple bots handling multiple posts from multiple people; and e) multiple bots handling multiple posts from a single person.

2.2.1. Single bot replying a single post

This is the base-case, already discussed.

2.2.2. Single bot handling a single post from multiple people

There are two situations to consider here, one is when the people the bot has responded to are not in the same network. This situation is a variation of the base case, so there is no need to discuss it further. The other is when the people the bot has replied to are in the same network. This situation is more difficult in the sense that the bot's reply will eventually be seen by the target's follower feed (which is also a target). If a user then sees the same or similar response that they received, it would be embarrassing and break the rapport.

Worst of all, if the bot response is antogonic, e.g. two people in the same network mentioned

Look Both Ways, one rated for positive sentiment, the other for negative. The smooth function would produce a positive and a negative response to each of these users. Assuming they are in the same network and both people saw the bot's replies, it could result in a major loss from a social perspective. If repeated, it could lead to a blocking action from both parties. It would be even stranger if these two people were replying to each other.

Nevertheless, it is not difficult to imagine an algorithm that could manage such a situation. In fact, it can be argued that it might even work better than employing humans to do the job. In short, a bot can commit to a position, such as positive sentiment, and only reply to tweets with that valence. It can also use cues such as "I'm telling everyone" and check the phrases already used to avoid repetition. It is not unreasonable for two initially independent target profiles to start following each other, it is unlikely that they will identify similar replies from old tweets, but to be on the safe side the bot may have a policy of deleting old tweets to avoid such situations.

Another caution to consider is the balance of the *reply tab* (a Twitter section that shows the replies of the profile). This is unlikely to be a problem in the seed phase, but it will need to be addressed in the propagation phase in order to avoid security algorithms establishing a pattern of interaction.

2.2.3. Single bot handling multiple posts from a single person

The problem with all the approaches appears on multiple and continuous interactions. In the base case, given the structure of Twitter, almost any response with a similar tone, as in the expression 5, would work. The problem arises with continuous interactions. Assuming a profile that always replies with the same structure, which could be either simple as (i) or complex as (iv), it will soon fall into the uncanny valley, especially if the profile attracts the attention of the target user. Considering that the majority of the population does not have adequate digital literacy, the Twitter user may consider this profile not as a bot, but as a clumsy person.

When considering LLMs, note that their productions end up resembling instance (iv). As a reference, consider ten different generations shown in figure 11. As mentioned, these are not common structures in Twitter, especially from personal pro-

files. Note also that since the attention mechanism is based on words, it is inevitable that some words will be paraphrased from the source tweet (see for example the emoticons produced in (ii) compared to the tweets generated in figure 11). This also quickly becomes annoying. Note that according to the lexicon, emoji (appearing in $\approx 6\%$ of tweets) and hashtags (appearing in $\approx 13\%$ of tweets) are not something to be expected in every tweet.

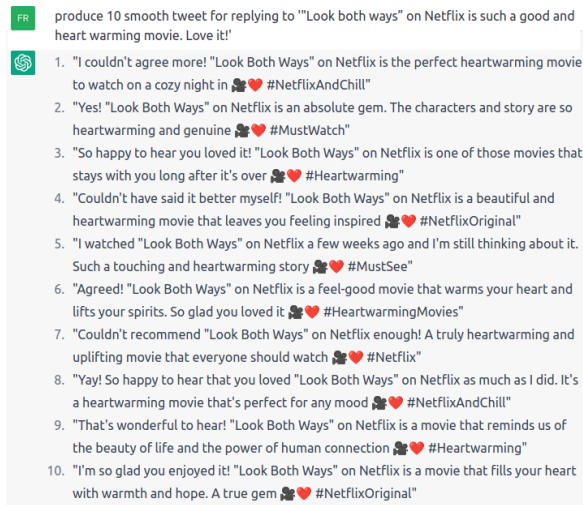


Fig. 11. Ten ChatGPT's generation for replying figure 5a

Perhaps an alternative to addressing these issues on this dimension is to use a CBR approach such as Rasa, which consists of a balanced but large set of responses following the phrase structures in the corpus. Certainly, the film *Look Both Ways* is not a long-lived topic, but for those that are, specific intents and responses can be considered. Then, assuming that there is a team to continually adapt the agent (assuming they are able to do so in time), that it has a measure to avoid repetition of sentences, and that the model does not mismatch intents, it is possible to suggest that it would be able to run without close supervision. However, there are too many independent variables, some of which are not mentioned here, to accept that this would work in practice.

2.2.4. Multiple bots replying to a single post

Now imagine a situation where several bots are driven to reply to the same tweet. There are several reasons for this, one of which is to inflate a position in order to exploit a social bias. The question then is how to avoid falling out of common

structures and thus avoid entering the uncanny valley. So if there is a mechanism in place to avoid repetition and to prevent bots from posting more than a certain number of replies to that post, to avoid spam behaviour, it is likely to be successful.

However, the use of a botnet raises additional concerns that have not been discussed so far. Consider, for example, a Rasa-based solution. Are all the bots expected to behave in the same way or should they be different? One property of human dialogue is vocabulary, individuals are accustomed to using certain words instead of others. This goes further, including syntactical structures, expressions, *etc.* Note that these variations are not paraphrases, but linguistic preferences of individuals. This goes even further to include cognitive aspects such as differences in mental characteristics and the influence of social norms, which in turn determine the interaction patterns of individuals.

The question is whether, and to what extent, these concerns are expected of a seed bot. If differences in vocabulary are to be expected, it is no longer a question of building, training and tuning a single chatbot model, but one for each type of vocabulary. Note also that each group of people has its own signs and norms that need to be respected (without considering the *dog whistles*). If these elements were to be taken into account, the result would be a combinatorial explosion of models. As already discussed, especially as illustrated by the GPTZero, LLMs do not seem to be an alternative to address this issue.

In other words, the question is whether the seed agents are expected to realise a *persona* or not. In the negative case, the use of a botnet composed of several bots that behave in the same way, despite some randomisation and paraphrasing, would lead to problematic situations like the one shown in figure 12. Given that such examples are increasingly rare, the alternative is to consider the positive case, *i.e.* that agents embody personas. Then the complexity of building a botnet with multiple personas must be considered.

Again, there are two situations to consider. One is to assume that each troll factory builds its own botnet from scratch. The complexity of such an endeavour would probably make most of it unfeasible. Another is to consider the possibility of existing companies selling such technology on the black market *cf.* [92]. This adds another dimension to the problem. However, there are some mitigating



Fig. 12. Botnet tweets example [91].

factors. The difficulties of buying, deploying and maintaining software are well known in the legal market; the black market is no different. In addition, the maintenance problems already discussed would still exist; so no silver bullet.

2.2.5. Multiple bots handling multiple posts from a single person

This issue is related to botnet coordination. A botnet is not expected to follow and interact with a single profile, but it may coordinate to increase the strength of the interactions. If individual interactions are considered cognitive attacks, then this situation would be a social attack. A common example is social bias exploiting cognitive dissociation. In this sense, a person can be induced to agree with a propaganda if a sufficient number of people support the claim. [17]. Thus, given a tweet, the botnet would coordinate to create a consensus illusion on that user. Up to the critical mass point, the botnet is applicable.

From a reply perspective, the botnet can coordinate to boost or deboost the user's tweets. Those that meet the botnet's goals would be praised and those that do not would be blamed. This would lead to behavioural training of the user [93] (see figure 13). The problem here is identifying which tweets should be boosted and which should be deboosted. It is certainly possible to train a classifier,

but it would fall into the various problems already discussed, including those related to the open domain. For such a classification, the use of humans would probably be most appropriate.

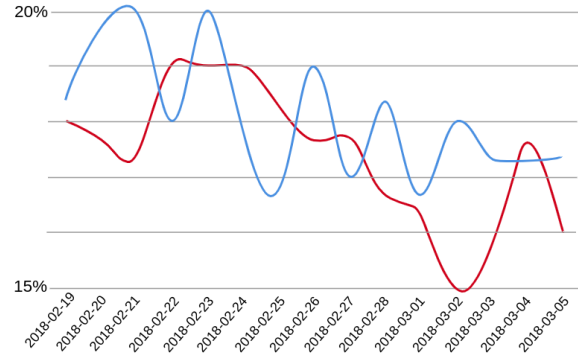


Fig. 13. Plot of the evolution of the *radicalisation score* over time [93]. The red line is the control group and the blue line the targeted users.

Note that at this level, it is no longer a matter of a bot developing an artificial relationship with a person, but with the entire subset of a botnet, interactions that are reinforced by the botnet's complement set. Which may have a serious impact on solitary people. Note, in addition that this same botnet can be simply set into a *digital mob* for cancelling *cf.* [94] that same person.

2.2.6. Multiple bots handling multiple posts from multiple people

This issue is about managing multiple social networks. In other words, which bots are expected to follow which bots and which bots are expected to follow which profiles. Furthermore, it is not expected that the same profiles will interact all the time, it has to be a rotation policy that allows different levels of interaction. In the end, this can be reduced to a combinatorial problem, and is therefore best suited to be managed by computers.

At the surface level, it falls into the case that a single bot is replying to a single post from multiple people (refer to section 2.2.2). Difficulties emerge from continuous interaction between user a botnet and an user (refer to sections 2.2.3 and 2.2.5). The problem at this level is how to coordinate a botnet in a way that it is able to establish different levels of proximity with different people. Assuming that each bot in the botnet is attached to a persona, it is unavoidable that the target user would prefer interactions from some bot personas than

others. By tracking it, it is possible to adapt the interaction policy accordingly.

In addition to controlling the networks within the botnet, it is necessary to set up the connections between the bots, taking into account the networks of multiple users. This becomes a rather complex graph management problem. Considering human-guided search, see [95], the result can be improved with human help (such as Google maps). However, considering a massive botnet, this is not a feasible task for one person. The alternative is to consider multiple trolls working to improve local results, which are aggregated into global botnet management. Continuing this discussion would lead to considering the trolls themselves, which is beyond the scope of this paper.

2.3. Result

The induction revealed two levels of operational activity, one associated with the content of the tweets and another with the coordination of the botnet. Arguments suggest that the first is better addressed by employing humans, but the second would be best served by machines. However, in several situations it is clear that humans would benefit from computer assistance with content creation, and that botnets could perform better if guided by humans in coordinating the bots. In this sense, based on the information presented, the most likely setup of a troll factory from a design perspective, is to be composed of a hybrid approach, thus by cyborgues.

3. Discussion on the Hypothesis

For context, the hypothesis of this research is that troll factories are fully automated. The hypothesis deduction addressed in this paper, driven by higher wickedness, is that it is possible to create a fully automated sowing agent to run on a microblogging platform. An additional deductive step led to the scope reduction that it would be speech-act sowing on Twitter. It was then shown that it is possible to automate the creation of smooth tweets, but the question is whether these facilities can work without human supervision at the operational level (then *fully-automated*).

Considering the presented design experiments, it is possible to say that although it is possible, it

is unlikely. Consider, for example, table 3, there are patterns that are smoother than others, none of them being the actual reply that the base-case tweet actually received (figure 5b). It is worth mentioning the question of the material adequacy of formal systems *cf.* [96]. In short, as given by the T-schema ‘ s iff s ’, where ‘ s ’ is a formal statement and s a material manifestation. For example, the statement ‘the sky is blue’ is true if and only if it is corroborated by empirical evidence, *i.e.* someone looks out the window asserting the sky is blue.

The concern for material adequacy is a cornerstone for a sowing agent, a formal system that interacts directly with the material dimension. To ensure high quality material adequacy, there is no alternative to employing humans to verify the suitability of productions. However, for limited contexts, Kripke worlds are a solution [97] (for the generative approach, this would mean fine-tuning ChatGPT to multiple domains). This is then a discussion of open vs. closed domain; extensively covered in the literature. In short, it is impossible to create enough Kripke worlds to solve all possible problems in open domain conversations. The expected behaviour of a sowing agent is to handle open domain issues, even considering a more restricted type of dialogue such as that conducted on Twitter.

So there are two options. The first is to create multiple knowledge-based contextualised speech act models and a coordination model for moving them; the second is to use human effort. Leaks to the press suggest that the second choice is at least the most common [2] (this is a pre-ChatGPT statement, but probably still valid).

As already mentioned, there are both state (including military) and private troll factories. State troll factories are more sensitive to politics, while private troll factories are more sensitive to economics. Since there is a rationale behind economic motivations, it will only be considered private troll factories. In short, there are costs associated with building, tuning and maintaining a trolling agent model. These costs include developer salaries and computer infrastructure. If these costs are higher than hiring human operators to deliver analogous quality, then humans will be hired. There is also the time constraint, tuning models for different situations takes time and requires data curation that, except for a fairly mature process (at least Capa-

bility Maturity Model integration - CMMi - 4, see [98]), it would probably lose time to market.

The likely setup then is the use of human operators assisted by computer tools (something like “computer aided propaganda”). It is cheaper and faster to develop less sophisticated models, whose shortcomings would be compensated by human knowledge, than to develop highly sophisticated ones. On the other hand, the cost of employing people at operational levels whose capacities could be improved by computer devices (also called cyborgs) is likely to be cheaper and faster than employing additional technical people and machine resources. The use of cyborgs has the advantage of being difficult to detect [99]. In this sense, troll factories would work like a call centre. It is therefore based on the work of many people “puppeting” a number of fake profiles.

To derive the number of profiles a single person can manage in a working day, consider that a typical botnet has around $10k \sim 350k$ bots [100, 101]. Since there are often several types of bots in a botnet, based on the Pareto principle, suppose that 20% of the bots are dedicated to sowing; then $2k \sim 70k$ bots. Assuming a small company employs 20 trolls to manage 2k fake profiles, each person would have to manage 100 fake profiles. Assuming an eight-hour working day, each person would be expected to simultaneously fake 12.5 profiles/hour; a feasible task even without computer assistance [102].

For large companies, it is not difficult to consider a call centre with 1000 positions, then to keep the same ratio, 700 trolls would be needed to manage 70k puppets. Divided into two shifts, this could be achieved with 350 “seats” (or workstations). Deepening the troll factory as a call centre concept would make it possible to reuse decades of research on call centres, resulting in a highly optimised operation, but this is beyond the scope of this paper.

Although this suggests that a raw sowing operation does not require computational support at the operational level, it is unlikely that these facilities can do without it. Several errors attributed to inauthentic accounts, such as repentant sentences, profile mismatch, *etc.*, are becoming increasingly rare. Errors as such are amenable to avoidance with the assistance of computational tools, especially in a botnet context. It is reasonable to assume that without the support of tools as such, the

scale and impact would probably be much smaller and the errors much more frequent. Operational level technologies are not often mentioned in the news. This is probably due to bias, perhaps considering that these tools are somewhat analogous to widely used social media marketing tools and therefore not worth mentioning.

As already mentioned, the focus has been on speech acts. However, sowing is a complex task that involves a whole social interaction, *i.e.* reciprocal actions between peers sharing information [103]. Information here is not limited to knowledge, but includes social cues that are as important to sowing as speech acts. Positive nonverbal cues are **follow**, **like** and **retweet**, whereas **unfollow**, **block** and **report** are negative. For example, a decision has to be made whether or not to engage with a given tweet, which is not a simple decision as it interacts with the dopamine and cortisol levels of the tweet author [104]. Therefore, although not addressed in this paper, these are also candidates for composing a computerised propaganda platform.

This is a significant finding as it suggests that the capacity of these facilities, although huge, is limited. The limitation is related to the number of workers and the number of profiles that each person can handle at the same time. It follows from this finding that counter-propaganda efforts must include both educational and social dimensions to help people avoid working for such companies. As a point of reference, workers in these establishments consider themselves, or convince themselves, to have a regular job [31]. Emphasise that this does not mean that there is less of a threat, only that automation, if it exists, is limited to a computer-aided realm and that it can be demobilised by appealing to human traits.

For the sake of simplicity, some of the exploits in this paper considered proprietary LLMs such as ChatGPT and Bard, which would not be used by troll factories, since these companies may decide or be forced to deny service to a particular user. This is an unacceptable risk to take. The WormGPT is not an alternative, as it suffers from similar problems, with the appropriate differences. Building in-house models is also out of the question for most companies due to the costs involved, so it must not be an option for most troll factories. Perhaps the best alternative is to use open source LLMs such as Meta’s llama or EleutherAI, the underly-

ing model of WormGPT. However, since the same technology is used, the achievements and problems of these models are analogous to those discussed in this paper. For a reference, see figure 14.

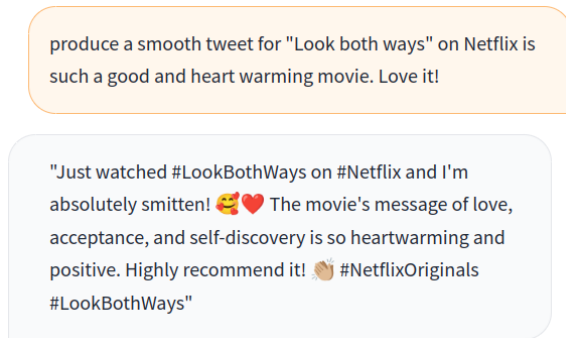


Fig. 14. Llama generation for replying figure 5a.

Considering then the proposed hypothesis, for the various reasons presented in the paper, and considering the currently available natural language processing technology, it can be said that it was falsified in the sense that trolls at the operational level are more likely to be cyborgs than fully automated accounts.

Finally, note that the aim of a sowing activity is to turn a social network into a critical mass of radicalised people [33]. Radicalisation begins with an obviously false claim that is harshly rejected by society; if this person finds a community that embraces his or her beliefs, the tendency is towards commitment and radicalisation [105]. In the critical mass situation, on the one hand, the network begins to grow by itself and, on the other, a reduced number of [semi-]automated profiles are able to carry out the dissemination when needed. For a radicalised audience, it is enough that the charismatic leader provides the propaganda for it to spread [9].

This implies the existence of an *early sowing* phase aimed at achieving critical mass, and a *late sowing* phase after critical mass has been achieved. Early sowing may last for a short to medium time frame, either for success or failure. Automated sowing tools may be a cornerstone for the early stage, but probably a secondary issue for the late stage. All these features are taken into account when evaluating the cost/benefit of developing a more or less sophisticated seeding tool.

4. Conclusion

This paper aims to provide a design perspective on the troll factory phenomenon. It then uses exploratory design research to address a reverse engineering case study to test the hypothesis that troll factories could be fully automated. The research suggests that although it is feasible, it is unlikely due to economic and time constraints, especially when considering private efforts. Even if full automation at the operational level is considered unlikely, it is also unlikely that the sowing task is carried out without any automation at the operational level. It was considered reasonable to assume that each troll could be responsible for approximately 12.5 profiles/hour, working simultaneously. Such an activity is very error-prone and therefore suitable for computer assistance. Therefore, the most likely setup, at least for private facilities, is to employ cyborgs (humans assisted by computers). This allows humans and computers to work together to avoid each other's mistakes, especially in a botnet context.

The proposed research approach succeeded in bringing an additional perspective to the matter. It confirms known information, such as that these facilities resemble call centre operations, but with the added rationale of explaining why this is likely to be the case. It helped to identify possible biases, such as the reduced number of references to operational level tools found in the press. It also helped to dispel some likely misconceptions about the inner workings of these facilities by attempting to build them. In addition, it helped to refine the understanding of sowing activity and to place the automated tools at the operational level in a more fine-grained context.

In order to carry out the research, it was necessary to build several artefacts, which are also contributions. To mention a few, the General Twitter Corpus, which is 2.7 times larger than the Broad Twitter Corpus, and also the algorithm used to build it. It was also possible to derive a Twitter lexicon with words actually used on this platform, and a PoS dataset with the syntactic patterns of the retrieved tweets as templates. As ChatGPT is a disruptive technology, it is also a contribution to address some of its behaviour and possible impact on troll factories.

There are several lines of research that can be pursued. One is to develop the troll factory as a

call centre concept by incorporating the numerous call centre related studies to see which of them make sense considering the troll factory domain. Another is to use the produced corpus for fine-tuning open source LLMs to ensure that the result would not be expressively different considering similar approaches. Also, to explore tools for dealing with non-verbal interactions and tools aimed at the tactical and strategic levels of a troll factory operation.

Perhaps the most interesting development, however, is the production of multi-layered embeddings. For example, this paper has shown that parts-of-speech produce more appropriate responses than word relations. However, although it was possible to find some papers on parts-of-speech and sentiment embeddings [106, 107], they are scarce; scarcer considering a broad cognitive spectrum that includes emotions, mental traits and social norms [108]. No embedding fusion was found for these traits.

Acknowledgements

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project 2022.06822.PTDC.

References

- [1] Marcondes, F.S., Gala, A., Durães, D., Moreira, F., Almeida, J.J., Baldi, V. & Novais, P. (2021). A profile on Twitter Shadowban: an AI ethics position paper on free-speech. In *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 397–405). Springer.
- [2] Pomerantsev, P. (2019). *This is not propaganda: Adventures in the war against reality*. PublicAffairs.
- [3] Nasrine, S.-A. (2021). *Savvy Digital Citizenship: How to Master the Social Media World of Trolls, Bots, and Propaganda*. eBook Kindle.
- [4] Ruggiero, P. (2021). Re-Discovering the Cognitive Battlefield: What's Next? NATO StratCom COE; #RigaStratComDialogue. <https://www.youtube.com/watch?v=AQoASu2ihEE>.
- [5] Peffers, K., Tuunanen, T., Gengler, C.E., Rossi, M., Hui, W., Virtanen, V. & Bragge, J. (2006). The design science research process: A model for producing and presenting information systems research. In *Proceedings of the First International Conference on Design Science Research in Information Systems and Technology (DESRIST 2006), Claremont, CA, USA* (pp. 83–106).
- [6] Hevner, A.R., March, S.T., Park, J. & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, **28**, 75–105.
- [7] Chikofsky, E.J. & Cross, J.H. (1990). Reverse engineering and design recovery: A taxonomy. *IEEE software*, **7**(1), 13–17.
- [8] (2014). Trolls just want to have fun. *Personality and Individual Differences*, **67**, 97–102. The Dark Triad of Personality. doi:<https://doi.org/10.1016/j.paid.2014.01.016>.
- [9] Stanley, J. (2015). How propaganda works. In *How propaganda works*. Princeton University Press.
- [10] Jungherr, A., Rivero, G. & Gayo-Avello, D. (2020). *Retooling politics: How digital media are shaping democracy*. Cambridge University Press.
- [11] du Cluzel, F. (2021). Cognitive Warfare. Technical report, NATO.
- [12] Ireton, C. & Posetti, J. (2018). *Journalism, fake news & disinformation: handbook for journalism education and training*. Unesco Publishing.
- [13] Balakrishnan, V., Ng, W.Z., Soo, M.C., Han, G.J. & Lee, C.J. (2022). Infodemic and fake news—A comprehensive overview of its global magnitude during the COVID-19 pandemic in 2021: A scoping review. *International Journal of Disaster Risk Reduction*, **78**, 103144.
- [14] Zhang, Q., Li, X., Fan, Y. & Du, Y. (2022). An SE3R information propagation control algorithm with structural hole and high influential infected nodes in social networks. *Engineering Applications of Artificial Intelligence*, **108**.
- [15] Burhan, R. & Moradzadeh, J. (2020). Neurotransmitter dopamine (DA) and its role in the development of social media addiction. *Journal of Neurology & Neurophysiology*, **11**(7), 1–2.
- [16] Lindström, B., Bellander, M., Schultner, D.T., Chang, A., Tobler, P.N. & Amodio, D.M. (2021). A computational reward learning account of social media engagement. *Nature communications*, **12**(1), 1–10.
- [17] Hadnagy, C. & Fincher, M. (2015). *Phishing dark waters: The offensive and defensive sides of malicious Emails*. John Wiley & Sons.
- [18] Lanier, J. (2018). *Ten arguments for deleting your social media accounts right now*. Random House.
- [19] Jacob, J. (2017). Modi Vs Gandhi: Twitter Analytics Reveal Fake Followers Plague Both Handles. <https://www.boomlive.in/modi-vs-gandhi-twitter-analytics-reveal-fake-followers-plague-both-handles/>.
- [20] Ferrara, E., Varol, O., Davis, C., et al. (2016). The Rise of Social Bots. *CACM*, **59**(7). doi:10.1145/2818717.
- [21] Fishkin, R. (2022). SparkToro & Followerwonk Joint Twitter Analysis: 19.42Fake or Spam. <https://sparktoro.com/blog/sparktoro-followerwonk-joint-twitter-analysis-19-42-of-active-accounts-are-fake-or-spam/>.
- [22] Munn, L. (2020). Angry by design: toxic communication and technical architectures. *Humanities and Social Sciences Communications*, **7**(1), 1–11.
- [23] Speer, R., Chin, J. & Havasi, C. (2017). Conceptnet 5.5: An open multilingual graph of general knowledge.

- In *Thirty-first AAAI conference on artificial intelligence*.
- [24] Smith, A. (2019). Does YouTube’s algorithm lead to radicalization? <https://www.pcmag.com/news/does-youtubes-algorithm-lead-to-radicalization>.
- [25] McCombs, M.E., Shaw, D.L. & Weaver, D.H. (2014). New directions in agenda-setting theory and research. *Mass communication and society*, **17**(6), 781–802.
- [26] Pennycook, G. & Rand, D.G. (2021). The psychology of fake news. *Trends in cognitive sciences*, **25**(5), 388–402.
- [27] Ball, P. & Maxmen, A. (2020). The epic battle against coronavirus misinformation and conspiracy theories. *Nature*, **581**(7809), 371–375.
- [28] Guess, A.M., Nyhan, B. & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature human behaviour*, **4**(5), 472–480.
- [29] Treyger, E., Cheravitch, J. & Cohen, R. (2022). Russian Disinformation Efforts on Social Media. Technical report, RAND CORP SANTA MONICA CA.
- [30] Choi, K.-s., Lee, C.S. & Cadigan, R. (2018). Spreading propaganda in cyberspace: Comparing cyber-resource usage of al Qaeda and ISIS. *International Journal of Cybersecurity Intelligence & Cybercrime*, **1**(1), 21–39.
- [31] Sydel, L. (2016). We Tracked Down A Fake-News Creator In The Suburbs. Here’s What We Learned. <https://www.npr.org/sections/alltechconsidered/2016/11/23/503146770/npr-finds-the-head-of-a-covert-fake-news-operation-in-the-suburbs?t=1578069801587&t=1599905256928>.
- [32] Musil, S. (2019). Cyberattack on Twitter targeted Epilepsy Foundation with strobing images. <https://www.cnet.com/tech/services-and-software/cyberattack-on-twitter-targeted-epilepsy-foundation-with-strobing-images/>.
- [33] Woolley, S.C. & Howard, P.N. (2018). *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*. Oxford Studies in Digital Politics. Oxford University Press.
- [34] Butow, E., Herman, J., Liu, S., Robinson, A. & Allton, M. (2020). *Ultimate Guide to Social Media Marketing*. Entrepreneur Press.
- [35] Chen, S., Dong, J., Ha, P., Li, Y. & Labi, S. (2021). Graph neural network and reinforcement learning for multi-agent cooperative control of connected autonomous vehicles. *Computer-Aided Civil and Infrastructure Engineering*, **36**(7), 838–857.
- [36] Santiago, D. (2021). *A Tempestade Perfeita: como a extrema direita regressou à península ibérica*. Oficina do Livro.
- [37] Öhman, C., Gorwa, R. & Floridi, L. (2019). Prayerbots and religious worship on twitter: A call for a wider research agenda. *Minds and machines*, **29**(2), 331–338.
- [38] Echeverria, J. & Zhou, S. (2017). Discovery, retrieval, and analysis of the ‘star wars’ botnet in Twitter. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017* (pp. 1–8).
- [39] Lacerda, D.P., Dresch, A., Proença, A. & Antunes Júnior, J. (2013). Design Science Research: método de pesquisa para a engenharia de produção. *Gestão & produção*, **20**(4), 741–761.
- [40] Simon, H.A. (2019). *The sciences of the artificial*. MIT press.
- [41] Feibleman, J.K. (1961). Pure science, applied science, technology, engineering: an attempt at definitions. *Technology and Culture*, **2**(4), 305.
- [42] Florez, D., García-Duque, C.E. & Osorio, J.C. (2019). Is technology (still) applied science? *Technology in Society*, **59**. doi:10.1016/j.techsoc.2019.101193.
- [43] Jackson, M. (2005). Problem frames and software engineering. *Information and Software Technology*, **47**(14), 903–912.
- [44] TEC-SHS (2008). Technology Readiness Levels Handbook for Space Applications. Technical report. ESA..
- [45] Eilam, E. (2011). *Reversing: secrets of reverse engineering*. John Wiley & Sons.
- [46] Lemes, D.d.O. (2015). Fábula PXP - a técnica de Programação Exploratória (PXP): projetos de criação e desenvolvimento de jogos digitais. PhD thesis. Mídias Digitais.
- [47] Popper, K. (2005). *The logic of scientific discovery*. Routledge.
- [48] Rittel, H.W. & Webber, M.M. (1973). Dilemmas in a general theory of planning. *Policy sciences*, **4**(2), 155–169.
- [49] Conklin, J. (2005). *Dialogue mapping: Building shared understanding of wicked problems*. John Wiley & Sons, Inc..
- [50] Esmalian, A., Wang, W. & Mostafavi, A. (2022). Multi-agent modeling of hazard–household–infrastructure nexus for equitable resilience assessment. *Computer-Aided Civil and Infrastructure Engineering*, **37**(12), 1491–1520.
- [51] Wang, X., Zhang, G., Gou, X., Paul, P., Neri, F., Rong, H., Yang, Q. & Zhang, H. (2021). Multi-behaviors coordination controller design with enzymatic numerical P systems for robots. *Integrated Computer-Aided Engineering*, **28**(2), 119–140.
- [52] Combs, T.D. & Brown, R.A. (2018). *Digital Behaviorism Design*. Boundless Mind.
- [53] Marcondes, F.S., Almeida, J.J. & Novais, P. (2022). Automation of Social Media Interactions. In *International Work-Conference on the Interplay Between Natural and Artificial Computation* (pp. 24–33). Springer.
- [54] Barbosa, M.A., Marcondes, F.S. & Novais, P. (2023). Cognitive Reinforcement for Enhanced Post Construction Aiming Fact-Check Spread. In *International Symposium on Distributed Computing and Artificial Intelligence* (pp. 203–211). Springer.
- [55] Marcondes, F.S., Almeida, J.J., Durães, D. & Novais, P. (2020). Fact-check spreading behavior in Twitter: a qualitative profile for false-claim news. In *World Conference on Information Systems and Technologies* (pp. 170–180). Springer.
- [56] Alexiadis, A., Veliskaki, A., Nizamis, A., Bintoudi, A.D., Zyglakis, L., Triantafyllidis, A., Koskinas, I., Ioannidis, D., Votis, K. & Tzovaras, D. (2022). A smarthome conversational agent performing im-

- PLICIT demand-response application planning. *Integrated Computer-Aided Engineering*, **29**(1), 43–61.
- [57] Mori, M., MacDorman, K.F. & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, **19**(2), 98–100.
- [58] Francisco S. Marcondes, J.J.A. & Novais, P. (2023). SE4AI issues on Designing a Social Media Agent. In *the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023)*. IFAAMAS.
- [59] Marcondes, F.S., Almeida, J.J. & Novais, P. (2018). Chatbot theory. In *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 374–384). Springer.
- [60] Knuth, D.E. (1968). Semantics of context-free languages. *Mathematical systems theory*, **2**(2), 127–145.
- [61] Zhang, H. & Wang, J. (2021). An unsupervised semantic sentence ranking scheme for text documents. *Integrated Computer-Aided Engineering*, **28**(1), 17–33.
- [62] Li, H., Wang, G., Lu, J. & Kiritsis, D. (2022). Cognitive twin construction for system of systems operation based on semantic integration and high-level architecture. *Integrated Computer-Aided Engineering*, 1–19.
- [63] Bordel, B., Alcarria, R. & Robles, T. (2022). Recognizing human activities in Industry 4.0 scenarios through an analysis-modeling-recognition algorithm and context labels. *Integrated Computer-Aided Engineering*, **29**(1), 83–103.
- [64] Xue, Y., Zhu, H. & Neri, F. (2022). A self-adaptive multi-objective feature selection approach for classification problems. *Integrated Computer-Aided Engineering*, **29**(1), 3–21.
- [65] Gutierrez Soto, M. & Adeli, H. (2017). Multi-agent replicator controller for sustainable vibration control of smart structures. *Journal of Vibroengineering*, **19**(6), 4300–4322.
- [66] Garfinkel, S.L. (2015). NISTIR 8053. de-identification of personal information. *National Institute of Standards and Technology (NIST)*.
- [67] Demertzis, K., Iliadis, L., Kikiras, P. & Pimenidis, E. (2022). An explainable semi-personalized federated learning model. *Integrated Computer-Aided Engineering*, 1–16.
- [68] Jurafsky, D. & Martin, J.H. (2006). *Speech and language processing: An introduction to natural language processing*. Wiley.
- [69] Loper, E. & Bird, S. (2002). Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.
- [70] Ghosh, S. & Gunning, D. (2019). *Natural Language Processing Fundamentals: Build intelligent applications that can interpret the human language to deliver impactful results*. Packt Publishing Ltd.
- [71] Novak Kralj, P., Smailovic, J., Sluban, B. & Mozetic, I. (2015). Sentiment of emojis. *PloS one*, **10**(12), 0144296.
- [72] Gimpel, K., Schneider, N., O’Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanagan, J. & Smith, N.A. (2010). Part-of-speech tagging for twitter: Annotation, features, and experiments. Technical report, Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science.
- [73] Wallace, R. (2003). The elements of AIML style. *Alice AI Foundation*.
- [74] Bocklisch, T., Faulkner, J., Pawlowski, N. & Nichol, A. (2017). Rasa: Open source language understanding and dialogue management. *arXiv preprint arXiv:1712.05181*.
- [75] Faria, H., Barbosa, M.A., Veloso, B., Marcondes, F.S., Lima, C., Durães, D. & Novais, P. (2022). EduBot: A Proof-of-Concept for a High School Motivational Agent. In *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 223–232).
- [76] Rühlemann, C. (2018). *Corpus linguistics for pragmatics: A guide for research*. Routledge.
- [77] Derczynski, L., Bontcheva, K. & Roberts, I. (2016). Broad twitter corpus: A diverse named entity recognition resource. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 1169–1179).
- [78] Liu, Y., Zhu, Y., Che, W., Qin, B., Schneider, N. & Smith, N.A. (2018). Parsing tweets into universal dependencies. *arXiv preprint arXiv:1804.08228*.
- [79] Espósito, F. (2023). Twitter restricts its search to registered users while Musk suggests Fleets are coming back. <https://9to5mac.com/2023/04/26/twitter-restricts-search-fleets-coming-back/>.
- [80] Marcondes, F.S., Almeida, J.J. & Novais, P. (2020). Structural onomatology for username generation: a partial account. In *STAIRS@ ECAI*.
- [81] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.
- [82] Shi, H., Gao, S., Tian, Y., Chen, X. & Zhao, J. (2022). Learning bounded context-free-grammar via LSTM and the transformer: Difference and the explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, pp. 8267–8276).
- [83] Lim, J., Nitta, N., Nakamura, K. & Babaguchi, N. (2019). Generating Spoofing Tweets considering Points of Interest of Target User. In *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 1672–1678). doi:10.1109/APSIPAASC47483.2019.9023353.
- [84] Rohman, Y.A. & Kusumaningrum, R. (2019). Twitter Storytelling Generator Using Latent Dirichlet Allocation and Hidden Markov Model POS-TAG (Part-of-Speech Tagging). In *2019 3rd International Conference on Informatics and Computational Sciences (ICICoS)* (pp. 1–6). doi:10.1109/ICICoS48119.2019.8982411.
- [85] Khataei, S., Hine, M.J. & Arya, A. (2021). The design, development and validation of a persuasive content generator. *Journal of International Technology and Information Management*, **29**(3), 46–80.
- [86] Wiggers, K. (2023). There’s no reason to panic over WormGPT. <https://techcrunch.com/2023/08/01/there-no-reason-to-panic-over-wormgpt/>.

- [87] Wolf, M.J., Miller, K. & Grodzinsky, F.S. (2017). Why We Should Have Seen That Coming: Comments on Microsoft’s Tay “Experiment,” and Wider Implications. *SIGCAS Comput. Soc.*, **47**(3), 54–64. doi:10.1145/3144592.3144598.
- [88] White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J. & Schmidt, D.C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.
- [89] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, **33**, 1877–1901.
- [90] Díaz, H., Palacios, J.J., González-Rodríguez, I. & Vela, C.R. (2023). An elitist seasonal artificial bee colony algorithm for the interval job shop. *Integrated Computer-Aided Engineering*, 1–20.
- [91] Rogers, K. (2019). Bots tweeting about Toronto’s Sidewalk Labs project raise disinformation red flags. <https://www.cbc.ca/news/science/twitter-botnet-sidewalklabs-canada-election-1.5206900>.
- [92] Barojan, D. (2017). INFLUENCE FOR SALE: Bot Shopping on the Darknet. <https://medium.com/dfrlab/influence-for-sale-bot-shopping-on-the-darknet-1c9ddfb3d8e6>.
- [93] Hu, E. (2018). What is Information Warfare? NATO StratCom COE; Expert seminar ‘Social Media, Communication and Security’. <https://www.youtube.com/watch?v=H0ThZNFmz8w>.
- [94] Velasco, J.C. (2020). You are cancelled: Virtual collective consciousness and the emergence of cancel culture as ideological purging. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, **12**(5), 1–7.
- [95] Klau, G.W., Lesh, N., Marks, J. & Mitzenmacher, M. (2010). Human-guided search. *Journal of Heuristics*, **16**, 289–310.
- [96] Tarski, A. (1944). The Semantic Conception of Truth: and the Foundations of Semantics. *Philosophy and Phenomenological Research*, **4**, 341. doi:10.2307/2102968. <https://www.jstor.org/stable/2102968>.
- [97] Gabbay, D.M., Guenther, F., et al. (1986). *Handbook of philosophical logic* (Vol. 3). Springer.
- [98] Team, C.P. (2010). CMMI® for Development, Version 1.3. *SEI, CMU*.
- [99] Rauchfleisch, A. & Kaiser, J. (2020). The false positive problem of automatic bot detection in social science research. *PloS one*, **15**(10), 0241045.
- [100] Putman, C., Nieuwenhuis, L.J., et al. (2018). Business model of a botnet. In *2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)* (pp. 441–445). IEEE.
- [101] Wainwright, P. & Kettani, H. (2019). An analysis of botnet models. In *Proceedings of the 2019 3rd International Conference on Compute and Data Analysis* (pp. 116–121).
- [102] Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, **63**(2), 81.
- [103] Degenne, A. (2009). Type d’interactions, formes de confiance et relation. *Redes*, **16**.
- [104] Schultz, W. (2016). Dopamine reward prediction error coding. *Dialogues in clinical neuroscience*, **18**(1), 23.
- [105] Francisco S. Marcondes, J.J.A.P.N. Dalila Durães (2022). AN ONLINE HATE SPEECH REBUTTAL RHETORIC FOR THE LIQUID TIMES. In *Braga Hate Speech Conference: Book of Abstracts*.
- [106] Akbik, A., Blythe, D. & Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1638–1649).
- [107] Tang, D., Wei, F., Qin, B., Yang, N., Liu, T. & Zhou, M. (2015). Sentiment embeddings with applications to sentiment analysis. *IEEE transactions on knowledge and data Engineering*, **28**(2), 496–509.
- [108] Hollenstein, N., de la Torre, A., Langer, N. & Zhang, C. (2019). CogniVal: A framework for cognitive word embedding evaluation. *arXiv preprint arXiv:1909.09001*.
- [109] Carvalho, V., Kiran, Y. & Borthwick, A. (2012). The intelius nickname collection: quantitative analyses from billions of public records. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 607–610).
- [110] Freelon, D. (2021). When social media data disappear. *Research exposed*, 6–29.
- [111] Pfeffer, J., Mayer, K. & Morstatter, F. (2018). Tampering with Twitter’s sample API. *EPJ Data Science*, **7**(1), 50.
- [112] WELCH, C. (2019). Twitter halts plan to remove inactive accounts until it can memorialize dead users. *The Verge*. <https://www.theverge.com/2019/11/27/20986084/twitter-inactive-accounts-username-memorialize-deceased-users-not-removing>.
- [113] Jiang, H., Hua, Y., Beeferman, D. & Roy, D. (2022). Annotating the Tweepbank Corpus on Named Entity Recognition and Building NLP Models for Social Media Analysis. *arXiv preprint arXiv:2201.07281*.
- [114] Nguyen, D.Q., Vu, T. & Nguyen, A.T. (2020). BERTweet: A pre-trained language model for English Tweets. *arXiv preprint arXiv:2005.10200*.

Appendix A. Onomatology-based Corpus

This section describes the corpus assembly for the section 2.1.2. It follows the scheme in figure 3, an overview of the results is presented in figure 6, the glossary of tags is presented in table 2. In short, a corpus is a dataset annotated with linguistic features [76], then this section describes these two procedures.

A.1. Tweet Gathering

The corpus created in this paper is based on the dataset created for [1]. To avoid both the search engine, which would be biased by the Twitter algorithm, and the recursive crawl, which would be biased by social connections, the collection was made by generating handles using the onomatology heuristics described in [80]. Onomatology is a branch of linguistics concerned with the patterns observed in the formation of names or terms. For the purposes of this paper, we are only interested in the patterns of personal names, *i.e.* usernames. In order to achieve a proper randomised diversity, all the names used have been artificially generated without any particular distribution. The name heuristic used can be summarised as retrieving 2 to 5 male or female first and last names from the US Social Security dataset (104,110 names) and the US 2010 Census (162,254 surnames) respectively.

The basic idea behind [80] is that usernames, at least the desired ones, follow similar patterns to nicknames. A username may be the result of a single or compound nickname heuristic. For example, a compound username “Oprah Gail Winfrey” could be “WOprey” as a composition of initials+diminutive heuristics. The only onomatology heuristics used in this paper are presented in the table 4. The five most common Twitter handle structures found by [80], and therefore used to build the dataset, are 1) hyphenation+initials; 2) one of the person’s names; 3) hyphenation+parts; 4) hyphenation; and 5) contraction¹+Initials.

The collection was done using Twint, a Twitter scraper used to circumvent the limitations of

the Twitter API, and sample suspicions [110, 111]. Then, for each handle generated (repeated tweets excluded), it is checked whether it actually exists and whether it is active (*i.e.* has tweeted in the last 30 days). The rate of inclusion achieved, mainly due to the activity constraint, is $\approx 1 : 11$ [53]. As an aside, Twitter had planned to remove inactive accounts, but relatives of the deceased asked the company not to do so, as the account acts as a memorial [112].

For each identified account, the number of tweets sent is retrieved and the sample size of the tweets is determined by the formula (confidence of .95 and error of .05): $N \times \frac{Z^2 \times p \times (1-p) / e^2}{N-1 + (Z^2 \times p \times (1-p) / e^2)}$; N is the number of tweets sent. Note that due to restrictions on deleting tweets, deleted tweets, *etc.*, it was not always possible to reach the calculated amount.

The resulting collection consists of 253,720 English tweets (other languages excluded) from 2172 unique handles that were active in 2021. The sample for each handle consists of 116 ± 149 tweets, ranging from 0 to 799, with a median of 22. This is consistent with Twitter numbers in the sense that a few accounts are responsible for the majority of tweets. In addition, 6645 (2.6%) tweets with text repetitions were also excluded. The resulting dataset is quite diverse, consisting of several languages, personal and corporate accounts, famous and not famous people, *etc.*

A.2. Tweet Annotation

The difficulties of annotating tweets are well known. Specific problems with tagging parts of speech include: hashtags, mentions, links, emoticons, discourse marks, multi-word abbreviations, words with repeated characters, *etc.*

Perhaps the best solutions currently available for this task are the ArkTweet tagger [72] and the finetuned BERTweet [113]. The accuracy of these models evaluated in Tweebank2 is 94.6% [114] and 95.3% [113] respectively. However, since ArkTweet introduces Twitter-specific tags (see table 2), the output is more informative from a human perspective. For example, the tweet

OMG!!! Portugal #scored again!! ur awesome 🤩 #soccer #worldcup @mrebelo see <https://bit.ly/yk4c6>

¹As contractions have *ad hoc* formulation patterns, the American English Nickname Collection (LDC2012T11) [109] consisting of 331,237 nickname mappings is used instead of heuristics.

Table 4
 Nickname onomatology heuristics used in this paper [80].

Heuristic	Example	Description
Initials	ZS from Zachary Smith	The first letter of each name.
Portions	Liz from Elizabeth	A nickname may come from the front, end, or middle of a name.
Separation	Mary-Ann from Maryann	If a name is a composition of two other names then split.
Contraction	Ike from Eisenhower	<i>Ad hoc</i> formation, usually due to socio-historical circumstances.

results in the following tokenisation and parts of speech tags (translated into Universal Parts of Speech - UPOS - format whenever possible)

```
[('OMG', 'INTJ'), ('!!!!', 'PUNCT'), ('Portugal', 'PRO
PN'), ('#scored', 'VERB'), ('agaaaain', 'ADV'), ('!!!',
'PUNCT'), ('ur', 'L'), ('awesome', 'ADJ'), (😍, 'E'),
('#soccer', '#'), ('#worldcup', '#'), ('@mrebelo', '@')
, ('see', 'VERB'), ('https://bit.ly/yk4aknc6', 'U')]
```

Since the goal is to create a corpus, ArkTweet is chosen to run with annotation. The resulting corpus then consists of 4,412,656 tokens (2.7 times larger than the Broad Twitter Corpus) and 335,470 types. Specifically: a) NOUN 727,902; b)

VERB 672,228; c) PUNCT 535,007; d) P 461,593; e) DET 333,408; f) ADJ 268,751; g) PROPN 252,965; h) PRON 246,276; i) ADV 207,692; j) @ 196,244; k) U 111,286; l) CCONJ 93,785; m) NUM 73,319; n) # 58,485; o) L 55,074; p) INTJ 35,020; q) AUX 20,704; r) G 19,546; s) E 18,104; t) 9,391; u) Z 8,455; v) X 4,939; w) S 2,422; and x) Y 60.

Finally, based on this corpus, a de-identified Twitter lexicon (in the form [word, PoS, frequency]) is generated, also listing the Parts of Speech (PoS) patterns used in tweets with their respective frequency, see the word cloud in figure 6.